Introduction to Statistics for Research

Survival Analysis

Martin Bland

Professor of Health Statistics

University of York

http://www-users.york.ac.uk/~mb55/

Survival, failure time, or time-to-event data:

- time from some event to death,
- time to metastasis or to local recurrence of a tumour,
- time to readmission to hospital,
- age at which breast-feeding ceased,
- time from infertility treatment to conception,
- time to healing of a wound.

The terminal event, death, conception, etc., is the **endpoint**.

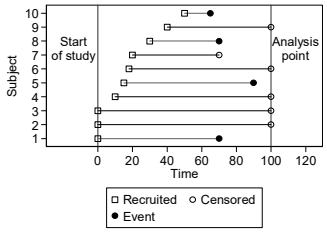
Often we do not know the exact survival times of all cases.

Some will still be surviving when we want to analyse the data.

When cases have entered the study at different times, some of the recent entrants may be surviving, but only have been observed for a short time. Their observed survival time may be less than those cases admitted early in the study and who have since died.

When we know some of the observations exactly, and only that others are greater than some value, we say that the data are **censored** or **withdrawn from follow-up**.

Recruitment, time to event, time to censoring:



Some censored times may be shorter than some times to events.

We overcome this difficulty by the construction of a life table.

Example

VenUS I: a randomised trial of two types of bandage for treating venous leg ulcers.

Treatments:

four layer bandage (4LB), elastic compression,
short-stretch bandage (SSB), inelastic compression.

Outcome:

time to healing (days).

VenUS I: SSB group, time to healing (days)

7 H	24 H	36 H	49 H	59 H	73 H	104 H	134 H
8 C	25 H	36 H	49 H	60 H	77 H	106 H	135 H
10 H	25 H	41 H	50 H	62 H	81 C	112 H	142 C
12 H	26 H	41 H	50 H	63 H	85 H	112 H	146 H
13 H	28 H	41 H	50 H	63 H	86 H	113 H	147 H
14 H	28 H	42 H	50 H	63 H	86 H	114 H	148 H
15 H	28 H	42 H	53 C	63 H	90 C	115 H	151 H
20 H	28 H	42 H	53 H	63 H	90 C	117 H	154 C
20 H	28 H	42 H	56 H	63 H	90 H	117 H	154 H
21 H	30 C	43 H	56 H	68 C	91 H	118 H	158 H
21 H	30 H	45 H	56 H	68 H	92 H	119 H	174 H
21 H	31 C	45 H	57 C	70 H	94 H	124 H	179 H
21 H	34 H	47 H	58 H	70 H	97 H	125 H	182 H
22 H	35 H	48 C	58 H	73 C	99 H	126 H	183 H
24 H	35 H	48 H	59 H	73 H	101 H	127 H	189 H
. .	. .	. .	. .	. .	. .	. .	. .

H = Healed C = Censored

VenUS I: SSB group, time to healing (days)

```
189 H  232 H  364 H  483 H  671 H
189 H  235 H  369 C  493 C  672 C
191 H  241 H  369 C  504 C  691 C
195 H  242 C  370 C  517 H  742 C
195 H  242 H  377 C  525 H  746 C
199 H  244 H  378 C  549 H  790 C
201 H  273 C  391 C  579 H  791 C
202 C  284 H  392 H  585 C  858 C
210 H  286 H  398 H  602 H  869 C
212 H  309 C  399 H  612 C  886 C
212 H  322 H  413 H  648 H  924 C
214 H  332 H  417 C  651 C  955 C
216 H  334 C  428 C  654 C
218 H  336 H  461 H  658 C
224 H  343 H  465 H  667 C

H = Healed   C = Censored
```

VenUS I: SSB group, time to healing (days), tabulated

```
t  C H   t  C H   t  C H   t  C H   t  C H   t  C H
 7 0 1   31 1 0   58 0 2   94 0 1  126 0 1  189 0 3
 8 1 0   34 0 1   59 0 2   97 0 1  127 0 1  191 0 1
10 0 1   35 0 2   60 0 1   99 0 1  134 0 1  195 0 2
12 0 1   36 0 2   62 0 1  101 0 1  135 0 1  199 0 1
13 0 1   41 0 3   63 0 6  104 0 1  142 1 0  201 0 1
14 0 1   42 0 4   68 1 1  106 0 1  146 0 1  202 1 0
15 0 1   43 0 1   70 0 2  112 0 2  147 0 1  210 0 1
20 0 2   45 0 2   73 1 2  113 0 1  148 0 1  212 0 2
21 0 4   47 0 1   77 0 1  114 0 1  151 0 1  214 0 1
22 0 1   48 1 1   81 1 0  115 0 1  154 1 1  216 0 1
24 0 2   49 0 2   85 0 1  117 0 2  158 0 1  218 0 1
25 0 2   50 0 4   86 0 2  118 0 1  174 0 1  224 0 1
26 0 1   53 1 1   90 2 1  119 0 1  179 0 1  232 0 1
28 0 5   56 0 3   91 0 1  124 0 1  182 0 1  235 0 1
30 1 1   57 1 0   92 0 1  125 0 1  183 0 1  241 0 1
```

VenUS I: SSB group, time to healing (days), tabulated

```
 t  C H   t  C H   t  C H   t  C H
242 1 1  378 1 0  549 0 1  790 1 0
244 0 1  391 1 0  579 0 1  791 1 0
273 1 0  392 0 1  585 1 0  858 1 0
284 0 1  398 0 1  602 0 1  869 1 0
286 0 1  399 0 1  612 1 0  886 1 0
309 1 0  413 0 1  648 0 1  924 1 0
322 0 1  417 1 0  651 1 0  955 1 0
332 0 1  428 1 0  654 1 0
334 1 0  461 0 1  658 1 0
336 0 1  465 0 1  667 1 0
343 0 1  483 0 1  671 0 1
364 0 1  493 1 0  672 1 0
369 2 0  504 1 0  691 1 0
370 1 0  517 0 1  742 1 0
377 1 0  525 0 1  746 1 0
```

The Kaplan Meier Survival Curve

t	C	H	n	d	s	p
0	0	0	192	0	192	192/192
7	0	1	192	1	191	191/192
8	1	0	191	0	191	191/191
10	0	1	190	1	189	189/190
12	0	1	189	1	188	188/189
13	0	1	188	1	187	187/188
14	0	1	187	1	186	186/187
15	0	1	186	1	185	185/186
20	0	2	185	2	183	183/185
21	0	4	183	4	179	179/183
22	0	1	179	1	178	178/179
24	0	2	178	2	176	176/178
25	0	2	176	2	174	174/176
26	0	1	174	1	173	173/174
28	0	5	173	5	168	168/173
30	1	1	168	0	168	168/168
.	.	.	.	.	.	.

n = number remaining

d = number of events

s = number surviving

p = proportion surviving

p = s/n

The Kaplan Meier Survival Curve

t	C	H	n	d	s	p
0	0	0	192	0	192	192/192 = 1.0000000
7	0	1	192	1	191	191/192 = 0.9947644
8	1	0	191	0	191	191/191 = 1.0000000
10	0	1	190	1	189	189/190 = 0.9947368
12	0	1	189	1	188	188/189 = 0.9947090
13	0	1	188	1	187	187/188 = 0.9946809
14	0	1	187	1	186	186/187 = 0.9946524
15	0	1	186	1	185	185/186 = 0.9946237
20	0	2	185	2	183	183/185 = 0.9891892
21	0	4	183	4	179	179/183 = 0.9781421
22	0	1	179	1	178	178/179 = 0.9944134
24	0	2	178	2	176	176/178 = 0.9887640
25	0	2	176	2	174	174/176 = 0.9886364
26	0	1	174	1	173	173/174 = 0.9942529
28	0	5	173	5	168	168/173 = 0.9710983
30	1	1	168	0	168	168/168 = 1.0000000
.	.	.	.	.	.	.

The Kaplan Meier Survival Curve

t	C	H	n	d	s	p	P
0	0	0	192	0	192	1.0000000	1.0000000
7	0	1	192	1	191	0.9947644	0.9947644
8	1	0	191	0	191	1.0000000	0.9947644
10	0	1	190	1	189	0.9947368	0.9895288
12	0	1	189	1	188	0.9947090	
13	0	1	188	1	187	0.9946809	
14	0	1	187	1	186	0.9946524	
15	0	1	186	1	185	0.9946237	
20	0	2	185	2	183	0.9891892	
21	0	4	183	4	179	0.9781421	
22	0	1	179	1	178	0.9944134	
24	0	2	178	2	176	0.9887640	
25	0	2	176	2	174	0.9886364	
26	0	1	174	1	173	0.9942529	
28	0	5	173	5	168	0.9710983	
30	1	1	168	0	168	1.0000000	
.	.	.	.	.	.	.	

Proportion surviving to time x:

$P_x = p_x P_{x-1}$

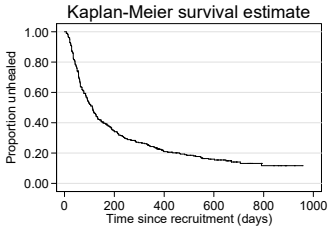
The Kaplan Meier Survival Curve

t	C	H	n	d	s	p	P
0	0	0	192	0	192	1.0000000	1.0000000
7	0	1	192	1	191	0.9947644	0.9947644
8	1	0	191	0	191	1.0000000	0.9947644
10	0	1	190	1	189	0.9947368	0.9895288
12	0	1	189	1	188	0.9947090	0.9842932
13	0	1	188	1	187	0.9946809	0.9790577
14	0	1	187	1	186	0.9946524	0.9738221
15	0	1	186	1	185	0.9946237	0.9685865
20	0	2	185	2	183	0.9891892	0.9581153
21	0	4	183	4	179	0.9781421	0.9371729
22	0	1	179	1	178	0.9944134	0.9319373
24	0	2	178	2	176	0.9887640	0.9214661
25	0	2	176	2	174	0.9886364	0.9109949
26	0	1	174	1	173	0.9942529	0.9057593
28	0	5	173	5	168	0.9710983	0.8795813
30	1	1	168	0	168	1.0000000	0.8795813
.	.	.	.	.	.	.	.

Proportion surviving to time x:

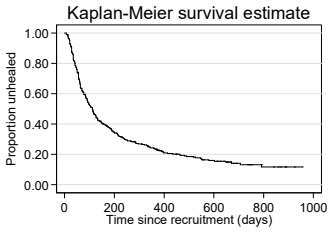
$P_x = p_x P_{x-1}$

The Kaplan Meier Survival Curve



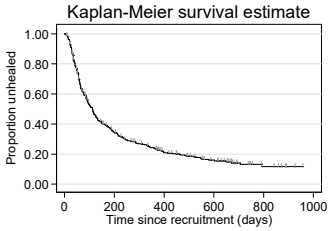
We usually present this graphically.

The Kaplan Meier Survival Curve



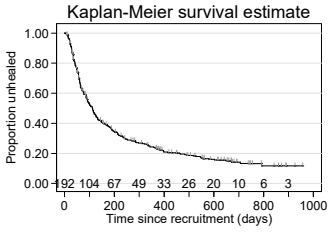
There is a step at each event. Steps get bigger at the number followed up gets smaller.

The Kaplan Meier Survival Curve



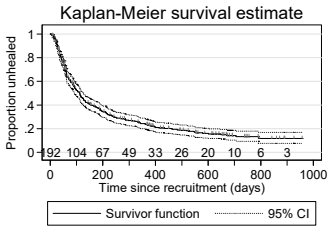
We often add ticks to indicate the censored observations.

The Kaplan Meier Survival Curve



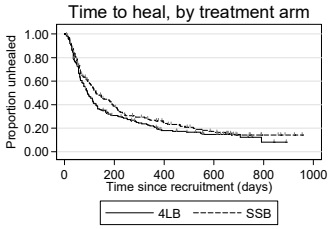
We can add the number remaining at risk along the bottom of the graph.

The Kaplan Meier Survival Curve



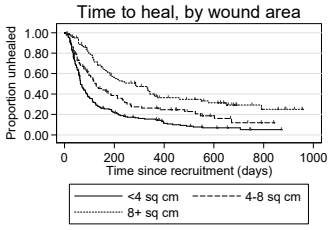
We can add a 95% confidence interval for the survival estimate. This is called the Greenwood interval.

The Kaplan Meier Survival Curve



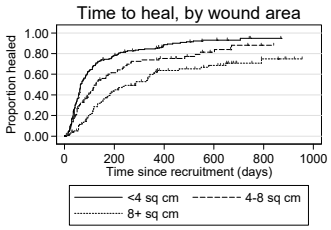
We can compare the two arms of the trial.

The Kaplan Meier Survival Curve



We can compare levels of a prognostic variable.

The Kaplan Meier Survival Curve



We can invert the graph and plot the proportion healed, called the **failure function** (opposite of survival).

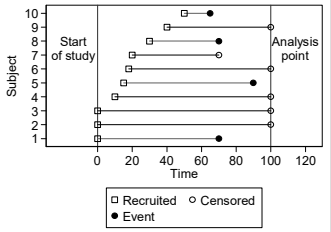
The Kaplan Meier Survival Curve

Assumptions

The risk of an event is the same for censored subjects as for non-censored subjects.

This means:

1. those lost to follow-up are not different from those followed-up to the analysis date,
2. no change in risk from start of recruitment to end.



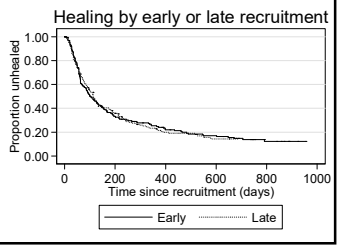
The Kaplan Meier Survival Curve

Assumptions

The risk of an event is the same for censored subjects as for non-censored subjects.

This means:

1. those lost to follow-up are not different from those followed-up to the analysis date,
2. no change in risk from start of recruitment to end.

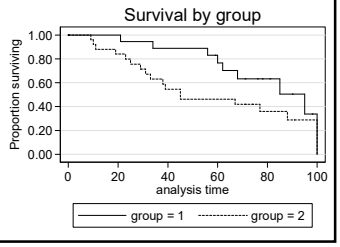


The logrank test

Greenwood standard errors and confidence intervals for the survival probabilities can be found, useful for estimates such as five year survival rate.

Not a good method for comparing survival curves. They do not include all the data and the comparison would depend on the time chosen.

Eventually, the curves will meet if we follow everyone to the event (e.g. death).



The logrank test

Survival curves can be compared by several significance tests, of which the best known is the **logrank** test.

This is a non-parametric test which makes use of the full survival data without making any assumption about the shape of the survival curve.

The logrank test

	SSB			4LB		
Time	n_1	c_1	d_1	n_2	c_2	d_2
0	192	0	0	195	1	0
7	192	0	1	194	0	3
8	191	1	0	191	0	0
10	190	0	1	191	0	0
11	189	0	0	191	1	0
12	189	0	1	190	0	0
13	188	0	1	190	0	1
14	187	0	1	189	0	3
15	186	0	1	186	0	1
17	185	0	0	185	0	1
20	185	0	2	184	0	2
21	183	0	4	182	1	4
.	.	.	.	.	.	.
.	.	.	.	.	.	.

Consider only times at which there is an event or a censoring.

n_1, n_2 = numbers at risk

c_1, c_2 = numbers of censorings

d_1, d_2 = numbers of events

The logrank test

	SSB			4LB			proportion with events
Time	n_1	c_1	d_1	n_2	c_2	d_2	$q_d = (d_1 + d_2)/(n_1 + n_2)$
0	192	0	0	195	1	0	0/(192+195)
7	192	0	1	194	0	3	4/(192+194)
8	191	1	0	191	0	0	0/(191+191)
10	190	0	1	191	0	0	1/(190+191)
11	189	0	0	191	1	0	0/(189+191)
12	189	0	1	190	0	0	1/(189+190)
13	188	0	1	190	0	1	2/(188+190)
14	187	0	1	189	0	3	4/(187+189)
15	186	0	1	186	0	1	2/(186+186)
17	185	0	0	185	0	1	1/(185+185)
20	185	0	2	184	0	2	4/(187+184)
21	183	0	4	182	1	4	8/(183+182)
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.

The logrank test

	SSB			4LB			expected events in group 1
Time	n_1	c_1	d_1	n_2	c_2	d_2	$e_1 = n_1 \times q_d$
0	192	0	0	195	1	0	192 × 0/(192+195)
7	192	0	1	194	0	3	192 × 4/(192+194)
8	191	1	0	191	0	0	191 × 0/(191+191)
10	190	0	1	191	0	0	190 × 1/(190+191)
11	189	0	0	191	1	0	189 × 0/(189+191)
12	189	0	1	190	0	0	189 × 1/(189+190)
13	188	0	1	190	0	1	188 × 2/(188+190)
14	187	0	1	189	0	3	187 × 4/(187+189)
15	186	0	1	186	0	1	186 × 2/(186+186)
17	185	0	0	185	0	1	185 × 1/(185+185)
20	185	0	2	184	0	2	185 × 4/(187+184)
21	183	0	4	182	1	4	183 × 8/(183+182)
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.

Sum e_1 to get expected events in group 1, SSB, = 160.57.

The logrank test

	SSB			4LB			expected events in group 2
Time	n_1	c_1	d_1	n_2	c_2	d_2	$e_2 = n_2 \times q_d$
0	192	0	0	195	1	0	195 × 0/(192+195)
7	192	0	1	194	0	3	194 × 4/(192+194)
8	191	1	0	191	0	0	191 × 0/(191+191)
10	190	0	1	191	0	0	191 × 1/(190+191)
11	189	0	0	191	1	0	191 × 0/(189+191)
12	189	0	1	190	0	0	190 × 1/(189+190)
13	188	0	1	190	0	1	190 × 2/(188+190)
14	187	0	1	189	0	3	189 × 4/(187+189)
15	186	0	1	186	0	1	186 × 2/(186+186)
17	185	0	0	185	0	1	185 × 1/(185+185)
20	185	0	2	184	0	2	184 × 4/(187+184)
21	183	0	4	182	1	4	182 × 8/(183+182)
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.

Sum e_2 to get expected events in group 2, 4LD, = 143.43.

The logrank test

Arm	Events observed	Events expected
4LB	157	143.43
SSB	147	160.57
Total	304	304.00

Apply the usual observed minus expected squared over expected formula:

$$\sum \frac{(O-E)^2}{E} = \frac{(147-160.57)^2}{160.57} + \frac{(157-143.43)^2}{143.43} = 2.46$$

This is from a chi-squared distribution with degrees of freedom = number of groups minus 1 = 2–1 = 1, P=0.1.

The logrank test

Can have more than two groups:

Area	Events observed	Events expected
<4 sq cm	176	122.24
4–8 sq cm	65	70.45
8+ sq cm	63	111.32
Total	304	304.00

chi2(2) = 46.84
P < 0.0001

Three groups, 2 df.

The logrank test

Assumptions

As for Kaplan-Meier.

1. the risk of an event is the same for censored subjects as for non-censored subjects,
2. survival is the same for early and late recruitment.

Test of significance only.

Misses complex differences where risk is higher in one group at beginning and higher in the other group at the end, e.g. the curves cross.

Cox regression

Also known as proportional hazards regression.

Sometimes we want to fit a regression type model to survival data.

We often have no suitable mathematical model of the way survival is related to time, i.e. the survival curve.

Solution: Cox regression using the proportional hazards model.

The **hazard** at a given time is the rate at which events (e.g. healing) happen. Hence the proportion of those people surviving who experience an event in a small time interval is the hazard at that time multiplied by the time in the interval.

The hazard depends on time in an unknown and usually complex way.

Cox regression

Assume that anything which affects the hazard does so by the same ratio at all times. Thus, something which doubles the risk of an endpoint on day one will also double the risk of an endpoint on day two, day three and so on. This is the **proportional hazards** model.

We define the **hazard ratio** for subjects with any chosen values for the predictor variables to be the hazard for those subjects divided by the hazard for subjects with all the predictor variables equal to zero.

Although the hazard depends on time we will assume that the hazard ratio does not. It depends only on the predictor variables, not on time.

The hazard ratio is the relative risk of an endpoint occurring at any given time.

Cox regression

In statistics, it is convenient to work with differences rather than ratios, so we take the logarithm of the ratio. This gives us the difference between the log hazard for the given levels of the predictor variables and the log hazard for the baseline, the hazard when all the predictor variables are zero.

We then set up a regression-like equation, where the log hazard ratio is predicted by the sum of each predictor variable multiplied by a coefficient.

This is Cox's proportional hazards model.

Unlike multiple regression, there is no constant term in this model, its place being taken by the baseline hazard.

Cox regression

In particular, we can estimate the hazard ratio for any given predictor variable.

This is the hazard ratio for the given level of the predictor variable, all the other predictors being at the baseline level.

Cox regression

Example: area of ulcer, a continuous measurement.

Coefficient (log hazard ratio) –0.0276
Standard error = 0.0064
Significance: z = –4.31, P < 0.001
95% confidence interval = –0.0402 to –0.0151

Hazard ratio = 0.973
95% confidence interval = 0.961 to 0.985.

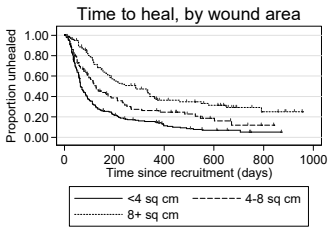
These are found by antilog of the estimates on the log scale.

This is the hazard ratio per sq cm increase in baseline ulcer area.

Bigger ulcers have lower risk, i.e. less chance, of healing.

Cox regression

Hazard ratio = 0.973, < 1.00. Bigger ulcers have lower risk, i.e. less chance, of healing.



Cox regression

Example: treatment arm.

Hazard ratio = 1.196
z = 1.56, P = 0.119
95% confidence interval = 0.955 to 1.498.

In this analysis SSB is the baseline treatment, so the risk of healing in the 4LB arm is between 0.955 and 1.498 times that in the SSB arm.

Compare logrank test: chi-squared = 2.46, d.f. = 1, P = 0.117.

The logrank test does not give quite the same P value as Cox regression.

Cox regression

Example: treatment arm.

Hazard ratio = 1.196
z = 1.56, P = 0.119
95% confidence interval = 0.955 to 1.498.

We can improve the estimate by including prognostic variables in the regression. Area is an obvious one:

	Haz. Ratio	z	P>|z|	95% Conf. Interval	
area	0.9723258	-4.35	0.000	0.960	0.985
arm	1.269221	2.07	0.038	1.013	1.590

Compare one factor hazard ratio = 1.196, P = 0.119,
95% confidence interval = 0.955 to 1.498.

Cox regression

Cox regression is described as semi-parametric: it is non-parametric for the shape of the survival curve, which requires no model, and parametric for the predicting variables, fitting an ordinary linear model.

The model is fitted by an iterative maximum likelihood method, like logistic regression.

Cox regression

Comparing models

We can compare nested models using a likelihood ratio chi squared statistic.

E.g. area only, LR chi-squared = 36.84, d.f. = 1

area + arm, LR chi-squared = 41.13, d.f. = 2

Difference = 41.13 – 36.84 = 4.29 with 2 – 1 = 1 degree of freedom, P = 0.038.

This enables us to test terms with more than one parameter.

Cox regression

Assumptions:

1. as for Kaplan Meier, the risk of an event is the same for censored subjects as for non-censored subjects,
2. the proportional hazards model applies,
3. there are sufficient data for the maximum likelihood fitting and large sample z tests and confidence intervals — rule of thumb at least 10 events per variable, preferably 20.

Cox regression

Checking the proportional hazards assumptions

There are several ways to do this.

We can look at the Kaplan Meier plots to see whether they look OK, e.g. do not cross.

Not very easy to see other than gross departures.

Cox regression

Checking the proportional hazards assumptions

There are several ways to do this.

We can look at the Kaplan Meier plots to see whether they look OK, e.g. do not cross.

Not very easy to see other than gross departures.

There are better plots, called log cumulative hazard plots, which we shall omit.